DOCUMENT RESUME

ED 163 057 TM 008 109

AUTHOR	Shepard, Lorrie A.
TITLE	A Checklist for Evaluating Large-Scale Assessment Programs. Paper #9 in Occasional Paper Series.
INSTITUTION	Western Michigan Univ., Kalamazoo. School of Education.
SPONS AGENCY	Carnegie Corp. of New York, N.Y.
PUB DATE	Apr 77
NOTE	60p.
AVAILABLE FROM	Mary Anne Bunda, The Evaluation Center, College of Education, Western Michigan University, Kalamazoo, Michigan 49008 ($2.00)
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	*Check Lists; *Educational Assessment; Evaluation; *Evaluation Criteria; *Evaluation Methods; *Program Evaluation; *Summative Evaluation
IDENTIFIERS	*Meta Evaluation

ABSTRACT

A checklist with five major categories is presented and discussed for its use in planning and carrying out an assessment program. Three other outlines are briefly presented for comparison: Stufflebeam's Meta-Evaluation Criteria; Scriven's Checklist for Evaluating Products, Producers, and Proposals; and Stake's Table of Contents for a Final Evaluation Report. Preparatory activities for assessment include staffing the evaluation team, defining the purpose, and identifying appropriate strategies for data collection. The first category of the checklist, goals and purposes, includes a number of different kinds of goals as well as criteria for judging goals. The technical aspects category includes several items about tests and also includes sampling, test administration, and reporting. There are subtopics under the heading of management such as planning and personnel. The category dealing with intended and unintended effects has two headings: people and groups who may be affected, and kinds of effects. The final category deals with dollar costs, costs in time, and possible negative effects of the assessment program. Finally, it is suggested that the results should be synthesized and contrasted with plausible alternatives. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED163057

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A Checklist for Evaluating Large-Scale Assessment Programs

by

Lorrie A. Shepard
Laboratory of Educational Research
University of Colorado
Boulder, Colorado 80309

April 1977

EVALUATION CENTER

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
M. A. Bunda
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

COLLEGE OF EDUCATION
WESTERN MICHIGAN UNIVERSITY
KALAMAZOO, MICHIGAN 49008

Paper # 9
in
Occasional Paper Series

IM008 109

WESTERN MICHIGAN UNIVERSITY
KALAMAZOO, MICHIGAN 49008

OCCASIONAL PAPER SERIES

The Occasional Paper Series is published by the Evaluation Center, College of Education, Western Michigan University. It provides information about new developments in evaluation to various educational audiences.

For information about previous publications in the series, see the inside of the back cover of this document.

All correspondence in reference to the Occasional Paper Series should be addressed to:

Information Officer
Evaluation Center
College of Education
Western Michigan University
Kalamazoo, Michigan 49008

EVALUATION CENTER

Daniel L. Stufflebeam
Director

Mary Anne Bunda

James R. Sanders

ABOUT THE CENTER

The Evaluation Center is committed to advancing the science and practice of evaluation. To serve its purpose the Center has developed components in research, development, instruction, dissemination, and service. Over the past ten years the staff of the Center has had significant experience in these areas working with school districts, state departments of education, foundations, and federal agencies.

The *Research Component* is designed to develop and test evaluation theory. Particular thrusts are testing the effects of school-based evaluation offices on school practice, and determining the special evaluation skill and knowledge profiles required by various groups in education. The *Development Component* is designed to work on ways and means of implementing evaluation theory. Included are techniques for meeting special evaluation objectives, new data collection instruments, and special procedures for communicating evaluative findings and relating them to practical problems. The *Instructional Component* is designed to provide in-service training in evaluation to various local, state, and national groups. While the Center is not a credit-granting agency, it collaborates with instructional units throughout Western Michigan University for the development and implementation of pre-service evaluation curricula. Emphasis in this work is at the doctors and masters levels. The Center serves as an instructional laboratory wherein students majoring and minoring in evaluation can gain practical evaluation experience. The Center is also committed to the development of novel and needed evaluation training materials and strategies. The *Dissemination Component* is designed to communicate research findings, new ideas, and materials to the education profession. *The Service Component* is designed to assist agencies in meeting their evaluation needs through two primary services. These are the design and implementation of evaluation studies and the development and installation of agency evaluation systems.

The staff of the Center has been involved in conceptualizing the CIPP Evaluation Model, researching the Item Sampling technique, developing the Advocate Team Technique, developing over 100 standardized tests, designing and conducting many in-service evaluation programs in evaluation. Other accomplishments include authoring the PDK book, EDUCATIONAL EVALUATION AND DECISION MAKING, and numerous journal articles and papers, designing and conducting many evaluation projects, and installing evaluation systems in schools, government agencies, and research agencies. Overall, the Center staff has directed externally funded projects in evaluation totaling more than four million dollars.

A Checklist for Evaluating Large-Scale Assessment Programs

by

Lorrie A. Shepard
Laboratory of Educational Research
University of Colorado
Boulder, Colorado 80309

April 1977

This paper was originally presented at the sixth annual Conference on Large-Scale Assessment, sponsored by the National Assessment of Educational Progress, Boulder, Colorado, June, 1976

There are many evaluation models around. The only justification for creating another is that it may make the evaluator's work easier or more insightful. The word model refers to an ideal plan, which, if followed, would result in the best possible evaluation. The connotation is that it will ensure that even a dull evaluator will do a good evaluation. Instead, it seems that existing models are either hopelessly elaborate or too general to be helpful in specific situations. Either of these deficiencies places more responsibility on the evaluator. If the model is too elaborate, the evaluator must decide which elements are essential and which can be safely omitted. If the model is over-simplified, the evaluator must be clever enough to put meaning on the barebones.

The authors of various evaluation models, (Alkin, 1969; Hammond, undated; Provus, 1969; Stake, 1967; Stufflebeam, 1971) have not told us exactly what to do but have stimulated our thinking about what evaluation ought to include. The purpose of this paper is this more modest goal. A checklist for evaluating large-scale assessment is offered to prompt evaluators to ask questions and seek effects in areas they might otherwise have missed. It should be more helpful than a general model because it is more specific.

In addition to improving the evaluation of assessment, a secondary purpose exists. Knowing what an evaluation of an assessment program ought to entail should be helpful to assessment staff. Not only could

---

I wish to thank my colleague, Dr. Gene Glass, for his careful reading of outlines and a draft of this paper. Special thanks are also due to Drs. Mary Anne Bunda and James R. Sanders of the Western Michigan faculty and to D. William Quinn of their staff for editorial and substantive suggestions.

they more effectively marshall their arguments and documentation, but the evaluation guidelines in this paper could foster self-study.

At the heart of this paper, after some preliminary discussion, is a checklist for evaluating assessment. A checklist is proposed rather than a model because a checklist is more down-to-earth and practical. A checklist can be used to make sure that all potential categories are considered but allows the evaluator to focus on those that are most salient for judging a specific assessment program.

Preview of the Assessment Checklist

The checklist for evaluating assessments has these major categories:

1. Goals and Purposes
2. Technical Aspects
3. Management
4. Intended and Unintended Effects
5. Costs

Full explication of these topics, with important subcategories and criteria, is given later in the paper. The complete checklist is presented in Table 4 on page 26. Briefly, the evaluation of assessment goals includes consideration of social significance and feasibility. Technical aspects include the content validity of tests, the appropriateness of sampling procedures, and the tailoring of reports for specific audiences. Management refers to the operations of the assessment; subcategories include sub-contractors and formative evaluation. The effects category is dicussed from a number of different aspects--not only what effects might be but how to look for them. Examples of effects would be increased teacher loathing of

tests or increased legislative funding for a new math curriculum. The cost category includes itemizing costs in dollars, hours, and attitudes. Ultimate judgments about costs and benefits are discussed in the final section of the paper following the checklist.

Sources for the Checklist

Although the usefulness of this paper will depend on the unique insights it provides for evaluating assessment, it would be foolish to start from scratch. Existing evaluation models are not immediately applicable to evaluating assessments; but, if properly translated, they already contain many of the important elements that one would wish to include in an evaluation of assessment. Almost any model, if studied in the assessment context, would add to our evaluation schema for assessment.

The purpose of this section of the paper is to review three existing evaluation frameworks: Stufflebeam's Meta-Evaluation Criteria (1974a), Scriven's Checklist for the Evaluation of Products, Procedures, and Proposals (1974), and Stake's Table of Contents for a Final Evaluation Report (1969). The purpose of this review is to acknowledge how much of the groundwork has already been laid for evaluating public service programs (see also, Sanders and Nafzinger, 1976) and to illustrate the additional steps needed to apply these schema to assessment.

Before discussing the applicability of these frameworks for the evaluation of assessment programs, one point should be made. These three sets of evaluation categories in Tables 1, 2, and 3 were devised by their authors for different purposes. None were intended for evaluation of assessment programs. There is some awkwardness in applying each to the evaluation of an assessment, since assessment is both the object of an evaluation and an evaluation activity itself. The Scriven and Stake outlines should be read as if assessment were the educational activity being evaluated. Stufflebeam's Meta-Evaluation criteria are for the evaluation of evaluation.

Stufflebeam's framework is unlike the other two in that it was not intended for judging ordinary educational endeavors. It was meant to be used to evaluate evaluations. Though it would be a grave mistake to equate evaluation and assessment, assessment is nonetheless an evaluation activity. Hence, the Stufflebeam criteria are generally applicable to the evaluation of assessment. These criteria are summarized in Table 1.

A creditable job of evaluating assessment could be done by following Stufflebeam's outline. Some of Stufflebeam's criteria could be used as they are. For example, Timeliness and Scope are subcategories in the Assessment Checklist and have essentially the same meaning as Stufflebeam gave them. The Internal Validity criteria, however, would have to include a number of assessment issues that are not immediately obvious, such as the content validity of assessment materials. A more serious difficulty in using only the Stufflebeam criteria is some apparent omissions. For example, the "utility" criteria neglect negative side-effects. This is not to suggest that

Table 1

Stufflebeam's Meta-Evaluation Criteria

I. Technical Adequacy Criteria

A. Internal Validity: Does the assessment[a] design unequivocally answer the questions it was intended to answer?
B. External Validity: Do the assessment results have the desired generalizability? Can the necessary extrapolations to other populations, other program conditions, and other times be safely made?
C. Reliability: Are the assessment data accurate and consistent?
D. Objectivity: Would other competent assessors agree on the conclusions of the assessment?

II. Utility Criteria

A. Relevance: Are the findings relevant to the audiences of the assessment?
B. Importance: Have the most important and significant of the potentially relevant data been included in the assessment?
C. Scope: Does the assessment information have adequate scope?
D. Credibility: Do the audiences view the assessment as valid and unbiased?
E. Timeliness: Are the results provided to the audiences when they are needed?
F. Pervasiveness: Are the results disseminated to all of the intended audiences?

III. Efficiency Criterion

Is the assessment cost-effective in achieving the assessment results?

[a]Note the substitution of assessment as the object of the Meta-Evaluation. Assessment is not evaluation, but it may be considered an evaluation activity.

Note: Table 1 derived from Stufflebeam, 1974(a).

Stufflebeam would overlook them in his own work; indeed, he has not when conducting actual evaluations (House, Rivers, and Stufflebeam, 1974). But a novice might make this error if he did not understand the full

implications of what was meant by "utility". Hopefully, then reducing such risks justifies developing a new plan for evaluation when other schemes already exist.

The Scriven Checklist (1974) in Table 2 has greatly influenced the content of this paper as well as the decision to adopt a checklist format. If conscientiously applied, the Scriven Checklist would result in a good evaluation of an assessment program. Once again, however, many of the important considerations are implicit rather than explicit. Scriven's Need category refers to the social importance of a program. The evaluator must determine real need not just audience receptivity. After all, "snake oil is salable" (Scriven, 1974, p. 13). Scriven's elaborations of the need criterion also includes the requirement of uniqueness. A need diminishes in importance if it is already served by other programs. Checkpoints for importance and uniqueness are subsumed in Assessment Checklist by evaluation of goals.

Scriven's Market checkpoint is distinguished from need. Some things like safety belts are needed but are not salable. Much of what is implied by Scriven's Market criterion corresponds to Stufflebeam's Disseminability. Elements of this issue emerge on the Assessment Checklist in both the Technical and Effects categories. Reporting is a technical matter but is also related to whether assessment results are used by their intended audiences.

Scriven suggests eight performance categories that, when combined with Market and Need, are the basis for judging the most impor-

Table 2

Scriven's Checklist for Evaluating Products, Producers, and Proposals

1.	Need (Justification) Need must be important and not served by other programs.
2.	Market (Disseminability) Many needed products are unsalable. Must be mechanisms for reaching intended market.
3.	Performance (True Field Trials) Try out the final product in typical setting with real users.
4.	Performance (True Consumer) Who are the real recipients? (Don't focus on teachers, legislators, and state department and miss parents, taxpayers, employers.)
5.	Performance (Critical Comparisons) Would another program be more successful or less costly (e.g., district-initiated testing programs)?
6.	Performance (Long-Term) Long-lasting effects.
7.	Performance (Side-Effects) Unintended effects.
8.	Performance (Process) What goes on when the product is implemented (e.g., teacher anxiety, student enjoyment, parent involvement)?
9.	Performance (Causation) Are the observed effects really due to the product? (Would the legislature have made those same appropriations anyway?)
10.	Performance (Statistical Significance) Are the effects considered in the above criteria real or due to sampling error?
11.	Performance (Educational Significance) Synthesis of items 1-10.
12.	Costs and Cost-Effectiveness Psychic as well as dollar costs.
13.	Extended Support Systematic continuing procedure for upgrading product.

Note: Table 2 derived from Scriven, 1974.

tant checkpoint, Educational Significance. In general, Scriven has focused on effects. There appears to be little concern for what Scriven (1967) earlier referred to as intrinsic or secondary evaluation, that is, judging the design of a program. Intrinsic evaluation in this context would include many of the technical criteria and some of the management considerations proposed in the Assessment Checklist. Instead, Scriven has focused on questions of program or product "payoff." Of course, technical adequacy is subsumed in the judgment of effects. If technical criteria are failed, effects will be near zero or negative. Certainly Scriven would not miss technical errors or management inefficiency. Nonetheless, there is some merit to making these requirements explicit so that others will not miss them.

Stake's (1969) Table of Contents for a Final Evaluation Report, Table 3, provides a larger perspective than is suggested by the other two schema. Rather than just proposing what kinds of information should be collected and judged, Stake has suggested some preliminary activities that ought to occupy the evaluators, e.g., determining the goals of the evaluation. One is reminded that if one were applying the Scriven Checklist or the Stufflebeam Criteria, there would be some additional activities required for beginning and ending the evaluation.

Much of the Stake outline is redundant with what has already been discussed. An evaluator using Scriven's performance checkpoints or Stake's outcome categories would focus on the same considerations. Stake does make explicit the identification of standards, although this can be inferred from Scriven as well. Stake's specification of the program section includes some descriptive elements

Table 3

Stake's Table of Contents for a Final Evaluation Report

I. Objectives of the Evaluation

- A. Audiences to be served by the evaluation (assessment staff, legislators, public).
- B. Decisions about the program, anticipated.
- C. Rationale, bias of evaluators.

II. Specification of the Program (Assessment)

- A. Educational Philosophy behind the program.
- B. Subject matter.
- C. Learning objectives, staff aims (assessment objectives).
- D. Instructional procedures, tactics, media.
- E. Students (students, teachers, district personnel who participate in the assessment).
- F. Instructional and community setting.
- G. Standards, bases for judging quality (in the assessment context those would be criteria for evaluating the assessment rather than performance standards on the instruments themselves).

III. Program Outcomes

- A. Opportunities, experiences provided.
- B. Student gains and losses (effects on the educational system).
- C. Side effects and bonuses.
- D. Costs of all kinds.

IV. Relationships and Indicators

- A. Congruences, real and intended (a report of the congruence between what was intended and what was actually observed).
- B. Contingencies, causes and effects.
- C. Trend lines, indicators, and comparisons.

V. Judgments of Worth

- A. Value of outcomes.
- B. Relevance of objectives to needs.
- C. Usefulness of evaluation information gathered (significance and implications of the findings).

Note: Table 3 derived from Stake, 1969.

which are not in the other frameworks and which have likewise been excluded from the Assessment Checklist and placed in a preparatory section. In the Relationships and Indicators portion, Stake goes further than Scriven in suggesting that effects be linked to program transactions.

In addition to the three general frameworks reviewed, existing evaluations of assessment programs were studied to glean specific criteria for judging assessments. The following reports were read, some in rough draft:

The 1974 site visit evaluation of National Assessment ("An evaluation of the NAEP," 1974);

The 1975 site visit evaluation of National Assessment ("An evaluation of the NAEP," 1975);

Greenbaum's study of NAEP (in press);

House, Rivers, and Stufflebeam's Assessment of the Michigan Accountability System (1974);

The Michigan State Department Response to the House et al, evaluation ("Staff response," 1974);

Stufflebeam's response to the response (1974 [b]);

Murphy and Cohen's article on the Michigan experience (1974);

A legislative evaluation of Minnesota's Assessment ("Minnesota," 1975);

The response of the Minnesota Assessment Staff ("Response to the senate," 1975);

A staff report to the Florida House Committee on Education ("Staff report to the committee," 1976);

Blue ribbon Panel report on Statewide Pupil Testing in New York State (in press);

Womer and Lehmann's evaluation of Oregon's Assessment (1974).

Complete references for these works are given in the bibliography.

Individual reports are referenced in the text when they provide substantiation or elaboration of a particular topic.

Preparatory Activities

In any evaluation there are activities which must precede judging merit. This section describes the following preparatory activities.

Staffing the Evaluation
Defining the Purpose of the Evaluation
Identifying Strategies for Data Collection
Describing the Assessment Program
Modifying the Checklist

Staffing the Evaluation

Evaluators ought to be knowledgeable about all aspects of an assessment. Glancing ahead to the checklist suggests many of the areas in which evaluators must be competent to make judgments. Evaluators should understand issues of test construction and sampling. Equally important is the ability of evaluators to judge social utility. Sixteen courses in measurement and statistics will not guarantee such insights.

Experts who possess all of the necessary knowledge are hard to find. A team approach may be used to ensure that various kinds of expertise are represented. I was a member of the 1975 site-visit team appointed by the National Center for Education Statistics to evaluate National Assessment. The membership of that team was strikingly diverse. Five of the nine members were not educationists. I was the only member of the team who could be identified with the traditional measurement-assessment community. Other members of the

team contributed perspectives and knowledge that I did not have. David Wallace from the University of Chicago and David Farrell of Harvard were experts in data analysis and computer utilization. Martin Frankel from the National Opinion Research Center played this same role but also poured over the details of NAEP's national sample and research operations; Frederick Hays from the Fund for the City of New York knew more about the management of large-scale bureaucracies than most academicians could ever hope to know. Maureen Webster from Syracuse Educational Policy Research Center and Norman Johnson of Carnegie-Mellon University had remarkable perspectives on national politics and could comment on which NAEP activities were likely to affect public policy and how.

Diversity has some drawbacks, of course. It's expensive to buy. In the NAEP example cited, some diversity might well have been traded for a deep study by a less varied team. Also, not having shared assumptions may make cooperation difficult. But this liability is also a strength since it protects against bias or narrow vision.

Though evaluators must be technically competent, there is a danger in selecting assessment experts to evaluate assessment. Assessors are likely to be philosophically committed to assessment. Others in the evaluation-measurement community have decided that testing is fundamentally harmful. It would be difficult to find many who were technically trained who did not already have leanings or biases in one direction or the other. These judges may be so committed to their beliefs about the value of assessments that they can

be counted on only to discriminate degrees of goodness or badness. Overall judgments of the social utility of assessment programs may more appropriately be left to judges who are less professionally and philosophically committed to a particular view on assessment.

A method of evaluation has little validity if the conclusions it generates depend on the personal bias of the evaluators conducting it. Care must be taken from the outset to ensure both the validity and credibility of the evaluation. A first step is to identify biases of potential evaluators. Ideally, we should be able to identify those who favor assessment and those who do not in the same way that Republicans are distinguished from Democrats or that the Republicanism of Goldwater is different from Rockefeller's. Some affiliations for or against assessment are public knowledge, some are subtle or unconscious and cannot be identified, but some are ones that the evaluators themselves would accede to. At the very least, candidates should be asked to rate themselves on the pro-assessment anti-assessment continuum. Then, an evaluation team should be composed to balance points of view. Evaluators of different persuasions might work together to reach consensus or might function as adversaries using a judicial or adversary approach. In his paper on Evaluation Bias and Its Control, Scriven (1975) recommended a "double-teaming" approach whereby two evaluators (or teams) work independently and then critique each other's reports.

There are many public and private complaints about the biases of the blue-ribbon panel of experts who evaluated the Michigan Account-

ability System (House, et al., 1974). Murphy and Cohen (1974) characterized the evaluators as follows.

> House has attacked the basic ideas behind accountability, and last summer he helped NEA develop its anti-accountability platform; Rivers has written about the evils of culturally biased tests; and Stufflebeam is an expert on evaluation and a stickler for research design. (p. 66)

Although they might disapprove of these one-line summaries, House and Rivers would probably locate themselves at the "skeptical-about-testing" end of the continuum. The characterization of Stufflebeam is most off the mark; he would probably say he was the pro-testing voice on the panel. One way to reduce these complaints, however, would have been to identify these philosophical differences in advance.

The evaluators had a contract with NEA and MEA to prohibit censorship of their report. But, they had no defense against the claim that they were hand-picked by the NEA to represent its point of view. One good protection would have been to invite the assessment staff or some other pro-assessment group (which the NEA is not) to nominate half of the evaluation team. My guess is that they would have selected Stufflebeam or someone very much like him, but such post hoc speculation is not as good as providing the guarantees at the outset.

Similar problems might arise if the Oregon State Department tried to quiet hostile anti-testers with the recent formative evaluation of their assessment. The evaluation was conducted by Frank Womer and Irv Lehmann, who are inveterate assessors. Their pro-assessment affiliation--as well as the fact that they conducted a formative evaluation and never meant to question the existence of the assessment--

would hardly be palatable to those who believe the assessment does more harm than good.

Detecting and controlling bias is not simple. Perhaps we need an inventory to administer to potential evaluators. It could test knowledge and attitudes including questions such as, "Are you more like Bob Ebel or Jim Popham?" It's the kind of inventory that would have to be validated using known groups. Items would be retained if they helped identify clusters such as Bob Ebel and Frank Womer, Ernie House and Bob Stake, or Wendell Rivers and Jane Mercer.

Another possibility is to select evaluators for their disinterest and train them in the technical issues. Scriven (1975) pointed out that the federal General Accounting Office is independent enough to be above suspicion though they may currently lack the expertise for conducting evaluation. In a similar context, he mentioned Alan Post, California's non-partisan Legislative Analyst. My colleague, Gene Glass, was recently an expert witness in a trial involving the interpretation of test results. He was impressed with the ability of two bright lawyers to grasp the technical issues and use the information after brief but intensive instruction. These examples make the idea of selecting evaluators who are not technically trained seem promising. It should certainly be considered for a large-scale evaluation.

Defining the Purpose of the Evaluation

Specifying the purposes of an evaluation makes it more likely that the purposes will be accomplished. This point has been discussed at length in the evaluation literature and will not be belabored here. The reader should beware, of course, that we now have to keep straight the goals of the evaluation, the goals of the assessment being evaluated, and the goals of the educational system being assessed.

The purposes of the evaluation will determine how the evaluation is conducted, what information is collected, and how it will be reported. Stake (1969) wrote that identifying goals of the evaluation involves recognizing audiences and the kinds of questions to be answered. One technique for accomplishing this is to simulate some possible conclusions and try them out on intended audiences. Would legislators want to know about recommended changes in test scoring or that other assessment programs obtain twice as much information for half as much money? Do classroom teachers want to know if the tests really measure what they are purported to measure or if the assessment should be made every five years instead of three?

One of the most important clarifications of purpose is the formative-summative distinction. Scriven (1967) originally coined the terms to distinguish between evaluations designed primarily to identify strengths and weaknesses for improvement (formative evaluation) and those intended to pass an overall judgment on a program (summative evaluation). Although Scriven (1974) subsequently stated that

"good formative evaluation involves giving the best possible simulation of a summative evaluation" (p. 9), the intention of the two types of evaluation is fundamentally different. Evaluators engaged in a summative evaluation are more likely to call into question the existence of the whole enterprise. Formative evaluators are more likely to assume the project's continuance and look for ways to correct weaknesses.

Identifying Strategies for Data Collection

Having decided on the evaluation's purpose, the evaluators must plan their work. Hopefully, the checklist explicated in this paper will serve to determine what kinds of information ought to be collected. There are, however, some over-arching issues about how information is gathered.

The well-known evaluation of the Michigan Accountability System (House, et al., 1974) was based on the examination of written documents and on testimony of various representatives of the educational system. Hearings were also used in New York [illegible], Millman, and Gordon (1974) to study the effects of [illegible] testing. Their findings were combined with results of a [illegible] admissions officers regarding the usefulness of [illegible] minations and a thorough analysis of testing costs. Two [illegible]tions of National Assessment have used site-visits. Greenbaum (in press) studied NAEP by reviewing transcripts of the founding conferences and by interviewing NAEP staff. In Florida, staff for the House Committee on Education ("Staff report," 1976) interviewed teachers and administrators from a random

sample of ten school districts to identify issues in education. Accountability and assessment was one of the areas probed.

Instructional programs are hard to evaluate, but there is still the possibility that with the right combination of tests and observations it will be possible to document program effects. The effects of assessment are more elusive. Evaluations of assessment will necessarily draw heavily on opinion surveys. This is risky. Opinions may be fairly accurate reflections of feelings about an assessment but may not be very good indicators of other kinds of effects. Somewhere in an educational research text was the observation that survey research is good for many things, but finding out which teaching method is best is not something one answers with a survey.

Evaluators will have to plan strategies so as not to miss effects. Some suggestions are given in the effects section of the checklist. Scriven's (1974) modus operandi method may be called for, searching for clues as Sherlock Holmes would have. Perhaps one could juxtapose release of assessment results and government spending or study the relationship between sending home pupil test scores and the incidence of unrequested parent visits to school. In the case of the Coleman report (1966), one would try to establish its effects on the funding of compensatory education by first interviewing politicians. Ask what facts they believe are "proven" or are "common knowledge" regarding equality of educational opportunity. Ask for sources they know of that substantiate these facts or ask their legislative aides.

Go to the materials they identify and seek their sources in turn. Although it oversimplifies the method, we could say that the more frequently Coleman appears in these bibliographies the greater the impact of that report on compensatory funding. A similar "search for connections" might be used to link the decline in SAT scores with the reinstatement of freshmen remedial writing courses.

This issue was not saved for the effects section of the paper since some decisions have to be made early on. Because opinions are likely to be an important source of information, evaluators will have to decide whom to talk to and when. Scriven (1974) argued convincingly that the evaluator ought to search for effects without knowing what the effects were supposed to be; thereby, his search would be less biased. Such goal-free evaluation might be particularly difficult in the assessment case since a number of persons will want to talk about goals, though this could be saved for last. Examining test materials and reports will not suffice since judgments about the adequacy of materials will depend on purposes. This point is given considerable attention later in the paper. A compromise which Scriven agrees to is that evaluators may begin goal-free and later switch to a goal-based approach. This would allow detection of unintended as well as intended effects. Then the evaluator could pursue program purposes and intents. To this end, the Assessment Checklist could be modified to consider effects first.

Each datum collected about an assessment program is fallible. The best procedure to ensure a fair evaluation is to cross-check each

opinion with other data sources. Our strategies should be much like that of investigative reporting now made famous by Woodward and Bernstein (1976). If a teacher testifies that all her children cried when they took the test, verify the effect. Don't just ask another teacher if her children cried--though a random sample of teachers might be appropriate. Check the story with representatives from other levels in the system. Interview children or watch while a test is given.

Evaluators should collect information by more than one method to balance the errors of a single method. In the now classic volume on Unobtrusive Measures, Webb et al. (1966) referred to this safeguard as"triangulation":

> ...The most persuasive evidence comes through a triangulation of measurement processes. If a proposition can survive the onslaught of a series of imperfect measures, with all their irrelevant error, confidence should be placed on it. (p. 3)

In addition, the evaluation design should be flexible enough to allow for the identification of additional data sources to corroborate the data collected in the first round.

Describing the Assessment Program

The Assessment Checklist is a set of categories for judging assessment programs. But, evaluation is not only judging. It has a large descriptive component. Some evaluation theorists have, at some times, argued that evaluation is only description--providing information to the decision-maker; they have argued that the applica-

tion of standards and values ought to be left to the reader (Stufflebeam, 1971; Stake, 1973). I have taken the larger view that the evaluator must provide judgments as well as facts. But, description is essential in either case.

Evaluation reports should include a narration about the following aspects of assessment:

History
Rationale and Purposes
Development
Procedures
Dissemination of Results

These descriptions will be needed by many readers as a context for the judgments that follow. Narration will also be important for better informed readers (like assessment staff) to establish that the evaluators knew what it was they were judging. Some disagreements about judgments may be traced to misinformation about operations or purposes.

If the evaluators have decided to begin their study goal-free, most of the descriptive information except procedures should be held until after the study of effects. But, in the final reports, context information should be presented first.

History includes: the testing programs and other educational indicators available before the assessment; preliminary committee work to create the assessment; legislation; implementation schedule; and the current stage of the assessment in the total chronology.

Rationale and purposes should include an account of why the assessment was instituted. An example of assessment purposes is this

excerpt from Florida Statutes (Section 229. 57 (2)) describing legislative intent:

> (a) To provide for the establishment of educational accountability in the public educational system of Florida;
> (b) To assure that education programs operated in the public schools of Florida lead to the attainment of established objectives for education;
> (c) To provide information for accurate analysis of the costs associated with public education programs; and
> (d) To provide information for analysis of the differential effectiveness of instructional programs.

Another example is taken from the Teachers' Manual for the California Second and Third Grade Reading Test:

> The purpose of the state testing legislation is to determine the effectiveness of education in California. The California Assessment Program is to provide the public, the legislature and school districts with information for evaluating the strengths and weaknesses of educational programs. State testing is not designed to meet classroom needs for individual pupil diagnosis. In order to provide reliable individual diagnostic information, tests for each pupil would have to be much longer than those used in this program; even with longer tests there still might be less than one-to-one correspondence between the objectives of the tests and the objectives of a particular school or district. Diagnostic testing is, therefore, a responsibility of the school district. (Teachers' Manual, 1974, p. 1)

Of course, there are purposes that are not acknowledged in assessment documents. Unofficial purposes will emerge as evaluators contact various assessment constituents.

Development of the assessment is the history of implementation: hiring assessment staff, identifying educational goals, specifying objectives, selecting subcontractors, test construction, field testing, and review panels. Many of the development activities will be judged by technical criteria in the checklist. Most evaluators will find it clearer to combine descriptions and judgments in their reports. In the process of their evaluation, however, thorough descriptions will precede judgment.

Procedures are the most observable aspects of the assessment. Procedures refers to the distribution of tests or proctors, children taking practice tests, teachers coding biographical data, children taking tests or performing exercises, and machine scoring of instruments.

Dissemination of results will be an important part in judging assessment usefulness. In the descriptive phase, "Dissemination of results" means identifying which kind of information is released to which audiences and in what form. What are the typical channels for dissemination of results after they leave the assessment office? Do principals pass information on to teachers? What is the frequency and tone of newspaper stories. (Incidentally, I recommend that state assessments institute a clipping service. Select a representative sample of 10 newspapers in the state and save any articles pertaining to assessment. Do not save only those articles that are examples of journalistic abuses. If such a file is not available, evaluators will have to do some archival work.)

Modifying the Checklist

In the introduction to this paper, I argued that evaluation models had to be "translated" to determine what they implied in a specific evaluation context. A method proposed specifically for evaluating assessment requires less translation. But, there are still unique evaluation purposes and features of the assessment being evaluated which require some special tailoring of the plan.

The best way to use the Assessment Checklist is to review the categories and specify what each means in a specific assessment context. Some subcategories may be omitted if they do not apply. However, these same cagetories may be reinstated if early data collection suggests they are relevant after all.

The categories of my checklist are not discrete. They interact and give each other meaning. This problem is present in the Scriven and Stufflebeam frameworks, but it is exacerbated by the effort here to be more specific. In the 1975 evaluation of NAEP, Maury Johnson ("An evaluation of the NAEP," 1975) outlined three sets of evaluation questions appropriate for evaluating any assessment program:

> ...mission, quality, and efficiency. The first set asks what the project is trying to accomplish in relation to what it could contribute to the Nation's well-being. The second examines how well current intentions are being realized, as reflected in the quantity and quality of products. The third set concerns the costs of these benefits and how well the project is run as an organization. Thus, the project can be viewed from a social and philosophical standpoint, from a technical standpoint, and from an economic and management standpoint. (p. 2)

These categories are comprehensive and distinct. The level of conceptualization is so synthesized that goals and attainments appear as one category. Still, these grand sets of questions imply most of the finer entries in the Assessment Checklist. The difficulty is that the more specific one becomes, the harder it is to separate the categories--order them, keep them from overlapping and interacting.

The most serious and pervasive interaction in the checklist is that of purposes with technical adequacy. Questions of sampling or test construction have to be considered in light of purposes or uses. Measurement texts always carry the adage that validity is not a property of a test, but of a use of a test. It is for this reason that goal-free evaluators may study effects without knowing purposes but must be informed about goals before they judge technical adequacy.

The checklist is arranged in an apparently sensible order. An evaluator may proceed in a different order or do several things at once. The best way to use the checklist is to become familar with the major categories so as to decide how these issues or combinations of issues ought to be addressed in a particular evaluation. It will serve its purpose well if an important area is probed that might otherwise be omitted.

Using the Checklist

Goals and Purposes

Kinds of Goals

Below are outlined possible purposes for an assessment program with illustrations of ways in which the assessment design should be modified to serve each purpose.

Pupil diagnosis is the most molecular purpose of testing. It means at least finding out what an individual child's strengths and weaknesses are. In a more specialized sense, it means finding out what kinds of errors are made so as to prescribe remediation. For this purpose, classroom teachers should not give tests to find out what they already know. For example, suppose we have an old-fashioned reading test of 50 items. If a child's teacher can predict accurately that the child will miss 20 and get 20 correct but is unsure about the remaining 10, the best use of assessment time (for this purpose) is to administer those 10 items--or better still 20 or 30 items of the same type. In this way, the teacher will acquire new information about what this child is able to do and may even have enough information to notice consistent errors.

Table 4

A Checklist for Evaluating Assessment

1. Goals and Purposes

Kinds of Goals	Criteria for Judging Goals
Pupil diagnosis	Importance
Pupil certification	Uniqueness
Program evaluation	Feasibility
program improvement	Scope
program selection	
Resource allocation	
Accountability	
Research	

2. Technical Aspects

- Tests
  - criterion-referenced vs. norm-referenced
  - content validity
  - cultural bias
  - empirical validity
  - reliability
- Sampling
- Administration of tests
- Reporting
  - data analysis
  - different reports for different audiences
  - interpretations

3. Management

- Planning
- Documentation of process
- Formative evaluation
- Personnel
- Support services
- Subcontractors
- Effective decision-making procedures
- Equal opportunity
- Redress of grievances
- Timeliness

4. Intended and Unintended Effects

People And Groups Who May Be Affected

- Students
- Teachers
- Parents
- Principals
- District superintendents
- District curriculum specialists
- Regional personnel
- State curriculum experts
- State legislators
- State governors
- Taxpayers
- Employers
- Special interest groups
- Researchers

Kinds of Effects

- Outcome and process effects
- Immediate and long-term effects
- Opinions
- Decisions
- Laws
- Resource allocation
- Effects of assessment technology

5. Costs

- Dollars
- Time
- Negative effects

The difficulty in serving this purpose along with others is that the 30-item test that one child should take is not the same as the one the next child should take.

Pupil certification is another general purpose for testing individual students. Tests used for this purpose must necessarily be the same for every individual and may produce results that are redundant to the classroom teacher. The redundacy is necessary to obtain the external verification of achievement.

The new California Proficiency Test, whereby 16 and 17-year-olds may obtain high school diplomas early, is an example of an "assessment" with this purpose. The New York State Regents' Examinations and the state administered bar exams are also examples.

Personnel evaluation is probably the most distrusted purpose for conducting a large-scale assessment. Judging teachers, principals, or superintendents on the basis of pupil test scores is invalid when there has been no correction for initial differences in pupil achievement, motivation, or ability to learn. Statistical adjustments for pupil differences are imperfect. While such adjustments may be acceptable for research purposes or for program evaluation, their potential for error must be taken into account when making individual decisions of merit. For a succinct refutation of the use of test scores to evaluate teachers, see Glass (1974).

When student performance data is to be used for personnel evaluation, class differences unrelated to the quality of instruction must be controlled either by random assignment of students to teachers or by statistically adjusting for differences in student ability. Even after this is done, confidence intervals should be constructed to reflect the accuracy in the statistical adjustments. One way to protect against over-interpreting chance differences is to use results from more than one year. If a teacher's student achievement (adjusted for ability) is very low year-after-year, and other teachers in the same circumstance have much greater achievement the evidence is compelling that the teacher is not effective.

Even after properly adjusted, it should be remembered that test scores only reflect attainment in a few of the many areas for which teachers are responsible.

Program evaluation is an appropriate use of assessment results. It should be noted, however, that the assessment results are not themselves the evaluation but may be used as data in an evaluation. Using assessment results to discriminate among programs requires that test content be program-fair and that tests have sufficient sensitivity to find important differences when they exist.

Program evaluation may be conducted at the national, state, district, or school building level. While there are many parallels in the requirements for conducting program evaluation at these different levels, they ought to be considered as separate purposes when judging technical adequacy or cost effectiveness.

Program evaluation includes both summative and formative purposes. For example, evaluation using assessment results may be used to select the best of three math curricula. It is more likely, however, that it will be used to identify strengths and weaknesses for use in modifying existing programs. Curriculum experts at both the state and local level may have the need for the most detailed reporting of assessment results. Developers of a reading program, for example, will need performance data which distinguishes between literal and interpretive comprehension and between phonetic and structural analysis skills.

Resource allocation is an often-stated purpose of assessment. Whether assessment results are actually used for this purpose or whether they should be is a tangled problem. Some state and federal funds are awarded to needy districts which have been identified on the basis of low test scores. Some politicians have objected that this system "rewards failure." As an alternative, Michigan adopted its Chapter Three incentives program. In the first year, funds were to be allocated on the basis of need, but in the succeeding years, money would be distributed on the basis of pupil progress. Districts with the highest

percentage of children who reached 75 percent of the month-for-month criterion would get the highest funding. The plan was to reward success instead of failure. Murphy and Cohen (1974) point out the fallacies in the plan and document the fact that it has never been implemented. House et al. (1974) objected to it because it oversimplified the problem of educating disadvantaged children and because it was demeaning to educators. I object to it because it ignores the political and moral justification for providing these funds in the first place. In order to "punish" the staff because of low pupil goals, funds are withheld; students who are in the need of the most help get the least assistance.

It is more defensible that assessment results would be used to establish state or national priorities: more dollars for education, fewer for highways or more dollars for math, fewer for science. Some differential allocation to districts may be possible if the above dilemma is resolved. In addition to allocating money, states may also distribute curriculum consultants on the basis of test results.

Accountability is implied in some of the other purposes (pupil certification and program evaluation), but it is also an end in itself. One connotation of accountability is the publicness of information. Satisfying the public's right-to-know may be a purpose of assessment.

This purpose requires that assessment results be both comprehensive and brief. The public and representatives of the public are impatient with mountains of data. They want simple answers.

Research may be a purpose of assessment. It is different from program evaluation in that it not only seeks the best instructional methods but tries to discern why one method is better than another. Research is often ridiculed as

an assessment purpose because it is not practical. It may be dysfunctional, but for a different reason. Research may be an awkward purpose for assessment because it requires tight controls that are not possible on so grand a scale.

Criteria for Judging Goals

Importance. The question here is whether the assessment is attempting to accomplish "good goals." One could hardly quibble with the importance of such goals as "improving education" or even "firing incompetent teachers," though the latter would have to be faulted on feasibility and technical criteria.

If assessments are hardly likely to aspire to bad goals, one might ask why this criterion is included. Well, relative judgments will have to be made in the ultimate analysis of cost and benefit. If a goal is good but not earth-shaking and costs a lot, it will not justify the assessment.

Uniqueness. Importance and uniqueness are the two ingredients in Scriven's Need checkpoint. I have separated these criteria because many goals are likely to pass the first and fail the second. The distinction prevents importance from camouflaging redundancy.

If an identified purpose is already served by other programs, it is less urgent. State assessment goals should receive low marks if they duplicate those of local district programs. NAEP ("An evaluation of the NAEP," 1975) received high marks for the goals which it served uniquely, adult assessment and administration performance exercises. Duplication of assessment efforts not only reduces the social need; it may also have direct deleterious effects in overtesting pupils.

Feasibility. Some assessment programs are guilty of inflated advertising. Many have promised that assessment will not only locate educational strengths and weaknesses, but will find solutions to ameliorate deficiencies. This criterion may sound like a judgment of technical adequacy. But it is not a matter of fine tuning the assessment design. Some intended purposes are beyond the scope of any assessment and should never have been promised. For example, finding and disseminating exemplary programs requires much more than assessment data and would involve product development and research.

Scope. Are the assessment goals too broad or too narrow? In the simplest sense, this may mean asking whether the assessment ought to include science and social studies as well as reading and math.

More significantly, however, this criterion pertains to the question about how many purposes can be combined before they begin to detract from one another. Curriculum experts and program evaluators need comprehensive information on content objectives. The California Reading Test for second and third graders includes almost 250 items (although each child only answers about 30 questions). Classroom teachers need information for every pupil. If they select different tests for each child, the results cannot be aggregated for accountability purposes or to answer the questions of the curriculum experts. If the comprehensive test was administered to every child, it would be extremely time consuming, give the classroom teacher a gread deal of redundant data along with the good stuff, and would have much more statistical power than was desired by either the curriculum experts or program evaluators. This dilemma is exemplified by Murphy's and Cohen's (1974) distress at the development of a new test in Michigan:

> The consequences boggle the mind. For example, a trial run of the first grade assessment (which had to be individually administered with these young children) wound up taking 30 hours per pupil just to test 50 minimum performance objectives. If a teacher started testing the first day of school, and tested all day, every day, until all 30 students had been individually assessed (assuming five hours of testing per day and an hour off for lunch in a nine-to-three school day), she would complete testing at 3:00 p.m. on the last day of the school year. (In its first year, the old assessment took only two hours.)
>
> This is ridiculous, of course, and the Department's testers realize it, but it does demonstrate some of the costs of being increasingly specific about objectives. In this case, the agency will probably either have to give up reporting all individual scores (by testing only samples on some objectives, which will make the results less useful to teachers), or it will have to test only high priority minimum objectives (as it did this year, which raises questions about what minimum objectives really are). (p. 68)

Some purposes will have to be excluded so that those remaining can be served well.

Technical Aspects

Tests

Criterion-referenced vs. norm-referenced. This consideration might be relegated to a discussion of test content or to the question of interpretation. The subject emerges so frequently as a topic of debate, however, that it deserves special attention. Criterion-referenced tests are preferred by many because they yield results that tell what individual children can do, rather than rank-ordering the test takers. Unfortunately, criterion-referenced testing may produce voluminous data that is ill-suited for some assessment purposes. Classroom teachers and curriculum makers are better served by specific information on 100 objectives but state legislators require more condensed information.

Norm-referenced tests are probably preferrable for placement purposes when the percentage of students who can be placed in certain situations is fixed. This is true of college admission. In other circumstances, knowing absolute,

rather than relative, performance levels is more diagnostic both for individuals and programs. Most criterion-referenced tests are, however, missing the "criteria" necessary for aggregate, social-policy interpretation. Most so-called criterion-referenced tests are content-referenced or objective-referenced but are missing the implied standards. Elsewhere, I have discussed the difficulties inherent in setting standards and have proposed some mechanisms for establishing criteria (Shepard, 1975). Without standards or comparative norms to assist in interpretation, the assessment results will be judged useless for certain audiences.

Content validity. The largest issues in content validity overlap with the question of scope in goal-evaluation. The content of the assessment should match the educational goals. There are tradeoffs allowed by other categories in the checklist to justify assessing in some goal areas and not others. It is much less defensible, however, to narrow the definition of those subjects selected for assessment.

Making sure that tests measure what they are supposed to measure is assured in part by proper development procedures. Deciding what to include or exclude in the definition of an assessment topic should not be left to closeted groups of subject matter experts. Parents and taxpayers also have something to say about what language or social studies tests ought to include. Greenbaum (in press) criticized NAEP for allowing the political clout of the school audience to disenfranchise lay representatives on objective-writing committees.

More detailed specification of test content is accomplished by writing objectives and items. Items should be reviewed to verify that they measure the intended objective and that, taken together, they provide a balanced representation of the content universe. There is a tendency to create many "easy to write" types of items and fewer addressing the objectives that are hard to measure. Someone has to keep an eye on the total picture.

Cultural bias. A test is culturally biased if certain children who have a skill are unable to demonstrate it because of extraneous elements clouding test questions. Webster, Millman, and Gordon (1974) helped clarify the distinctions between those aspects of culture that are relevant to achievement and those that are irrelevant and indicative of bias: "A test of ability to read road signs printed in English would have a strong and legitimate culture-specific content, while a test of reading comprehension might be biased by a culture-specific set of passages that comprise the test." (p. 16)

Tests are harmful if they were normed on majority groups (or groups in which the majority dominated) and are used to make irreversible decisions about individual children. Using test results to group children for instructional purposes is not harmful, however, if there is easy and frequent movement in and out of groups. When tests are used to evaluate programs, there is less concern that individual children will be diagnosed incorrectly, but it is still essential that each child be able to "do their best."

Empirical Validity. Tests are valid if they measure what they purport to measure. Content validity, discussed above, is established by logical analysis of test items referenced to the intended content universe. Additional evidence is needed, however, to verify the soundness of both the test-construction logic and the content analysis. Empirical evidence of test validity is typically obtained by correlating test scores with other measures of the same skills. Since these various criterion measures are also fallible, it is preferrable to use more than one validation criterion. If a new test is developed for an

assessment program, it may be advisable to study the agreement between student scores and teacher ratings of student skills and to examine correlations between subtests and other more lengthy tests available for assessing the same subskills. For an excellent summary of validation procedures see Cronbach (1971).

As was stated earlier, validity is not inherent in a test but depends on how a test is used. The uses that will be made of test results must, therefore, be taken into account when seeking evidence of validity. For example, when tests are to be used for program evaluation, evidence is needed of sensitivity to instruction or sensitivity to between-program differences (Airasian and Madaus, 1976; Rakow, Airasian & Madaus, 1976). If tests are to be used to select individuals, predictive validity should be well established. The evaluator or evaluation team are not responsible for collecting such data but must certainly judge the adequacy of existing validity data.

My consideration of validity here is necessarily brief because there are so many items in the checklist to cover and because the importance and method of test validation are known to both assessors and evaluators. I have introduced a few special issues to think about in the context of large-scale assessment. Such an abbreviated treatment should by no means be mistaken for an underrating of the importance of test validity. If the assessment instruments have serious flaws, the entire assessment effort is invalid. If the tests are not content valid, are culturally biased, or lack evidence of empirical validity, then the assessment results are questionable.

Reliability. Reliability is a prerequisite for validity. In order to measure accurately, an assessment instrument must measure dependably. Traditionally, reliability is best operationalized by test-retest correlation

coefficients. However, when many children receive near-perfect scores, which may be the case when tests are used to certify minimum competencies or to discriminate among programs rather than children, correlations will be near zero because of restriction of range. Additional information will be needed to make a sensible interpretation. In some instances, it would be appropriate to statistically correct for range restriction. Or, depending again on the use of the tests, it may be appropriate to examine the stability of program differences rather than pupil differences. Some evidence must exist, however, that assessment results are not quixotic and that conclusions would not be altered if children were tested on another day or in another setting. Articles by Huynh (1976) and Subkoviak (1976) indicate some of the progress being made at developing appropriate reliability statistics for criterion-referenced or mastery tests.

Sampling

Sampling must be appraised in light of assessment purposes. If the Michigan first-grade test (cited earlier) was ever implemented statewide, it would be a disaster due to the divergent purposes of individual testing and state-level testing. To meet the state level purposes efficiently, sampling is prescribed. For classroom teachers' use, every-pupil testing is required.

If sampling is appropriate, many more technical issues remain: adequate specification of the population, practical sampling methods (e.g., cluster sampling vs. random sampling), correct representation of sub-populations for reporting purposes, efficient sampling (small standard errors), and adequate follow-up of non-respondents.

Administration of Tests

Judging the administration of the assessment hinges on two major questions: do testing procedures yield the most accurate results possible and is the disruption as little as possible?

Test administration includes the selection and training of testers and the clarity of instructions to teachers. In most instances, evaluators should also inspect instructions to district superintendents and building principals to see if they have been well informed about the assessment being conducted in their jurisdiction. Most importantly, evaluators will have to judge how appropriate test format and vocabulary are for the children to whom it is administered. An assessment would get low ratings in this category if separate answer sheets were required for children in the primary grades. The California Assessment Program would get bonus points because every test is accompanied by a practice test to make sure that all children know how to respond. At least some practice questions are essential for all ages of respondents.

Reporting

Reporting is essential to make an assessment useful. Evaluation in this area is likely to require study of effects as well as study of assessment endeavors.

Data analysis. Routine data processing will have to be judged by the often incompatible criteria of accuracy and efficiency. Data analysis is more complicated. It requires judgments about statistical correctness but also prompts the larger question about appropriateness of selected variables. This may turn out to be a major issue in evaluating the utility of the assessment. Some critics of NAEP, for example, have insisted that results should be reported for Chicanos or for Spanish-speaking respondents. There are fairly large cost considerations, of course, that must be weighed against the gain in information.

Different reports for different audiences. Basically, what we are looking for are reports that contain the right kind of information and are understandable to their respective audiences.

Recommendations for better reporting practices were made in Shepard (1975). Intrinsic evaluation of reporting documents involves consideration of length, medium selected for presentation, availability of personal explanations, journalistic style, use of visual displays, and avoidance of technical jargon. Empirical verification of the adequacy of reports should come from true field trials or from follow-up studies of reporting in the full-scale assessment.

Interpretations. Some assessment exercises have self-evident meaning. It is clearly bad, for example, that in the first citizenship cycle, NAEP found that 25 percent of the nation's 17 year olds believed that they had to vote according to their party registration. But most assessment results require additional information to give them meaning. The extra ingredient must either be a performance standard or norm against which the obtained results may be compared.

Some assessors are fearful of making interpretations because they believe it will undermine their neutrality, an attribute considered essential for working cooperatively with schools. However, assessors could take responsibility for seeing that interpretations are made without passing judgment themselves. A number of states have begun the practice of inviting subject-matter experts to review results and publish their interpretations. This practice will ultimately have to be judged by its effects, but, in general, should be positively rated by the evaluators. As a word of caution, such responsibility for interpretation should not be left only to the school people. Perhaps the League of Women Voters or the NAACP should be invited

to make interpretations as well. Greenbaum (in press) proposed an additional group of reactors, political and social scientists. In a similar vein, I heard in a committee meeting a reference to the ultimate reporting vehicle: "Your results probably won't receive public notice until somebody famous writes a book about you."

Management

Checkpoint three calls for judgments of facilitating or enabling activities. Only two of the subpoints, formative evaluation and grievance procedures, are ends in themselves. The remaining topics are considered because they are likely to affect program outcomes.

Planning. I have a personal aversion for elaborate PERT charts, but it is essential that the assessment staff identify each step in development and implementation and plan who will work on each activity during which weeks and months. There is a tendency in all of us to work methodically at first and then run out of time for the last segments of a project. This paper had more attention at the beginning than at the end. Perhaps its only salvation is that the total outline was decided before the detailed work began.

Assessment staffs ought to be scrutinized particularly for how tney deal with unreasonable expectations. If they are being asked to do more than is possible, do they have good "coping behaviors"? Are they able to set priorities? Are they able to petition effectively for more resources or diminish expectations ("This will take two years instead of one.")?

Documentation of process. This is a record keeping requirement. Work may be duplicated if good records are not kept. Compromises reached about

test content or sampling plans should be well documented, specifying the arguments on each side and the details of the compromise. Faulty memory on these matters will require that consensus be constantly reestablished. Though constant debate would be dysfunctional for the operation of the assessment, new compromises will have to be reached from time to time. In these instances, an accurate history would make the deliberations more efficient.

Documentation may be maintained by recording decision-making meetings or by keeping a file of working papers summarizing arguments and agreements.

Formative evaluation. Scriven's Extended Support checkpoint causes us to look for mechanisms whereby the product will be continually upgraded. Formative evaluation may be a comprehensive study of the assessment or it may be a combination of separate activities such as field testing instruments and reports. It should, in either case, address questions of social utility. Is the assessment being used? Are there unintended uses that can be expanded? Scriven (1974) asks the question, Are there others who are producing a better product to serve the same need? "One decision that should remain open is the decision to cease production, even if it is commercially profitable to continue, when the evidence clearly indicates the existence of a superior product that can reasonably be expected to take over a free market" (p. 21).

The assessment program should be graded on the provision for and use of formative evaluation.

Personnel. The next three categories, Personnel, Support Services, and Subcontractors, should be taken together to ensure that all the necessary competencies are available to conduct the assessment.

Assessment requires political insight, technical expertise, management skills, writing ability, and a good sense of humor. Each member of the assessment team does not have to have all of the skills, but they must all be represented. An assessment run by experienced educators with lots of political savvy is just as imbalanced as one run by a dozen statisticians. If some of the necessary skills are missing in the assessment staff, assessment personnel have to be clever enough to know when to hire consultants or subcontractors.

Support services. At the simplest level, this item means asking whether work is being thwarted because six staff members share one secretary. Evaluators will not want to waste their time measuring office space and counting filing cabinets, but gross inadequacies or excesses in materials or services should be challenged. If assessment staff are housed on separate floors so that conferences are difficult to arrange, evaluators may give this ailment special attention. If computer facilities are grossly underused, time-sharing arrangements with other agencies might be recommended. Consultants should be used when needed to augment staff expertise.

Subcontractors. Writing RFP's (request for proposals) is an art that assessment staff will be graded on. Assessment staff are responsible for deciding which task can most appropriately be assigned to a subcontractor, for selecting the subcontractor, and for carefully monitoring subcontract activities. For example, a minor criticism of NAEP ("An evaluation of the NAEP," 1975) was that they had "farmed-out" a research report on background variables. At issue was not the quality of the report produced by the subcontractor, but the opportunity lost to the NAEP staff. It's bad for morale to

religate staff to routine tasks and give more inspired assignments to a subcontractor. In addition, NAEP staff might lack the familiarity with the intricacies of the report that they could have had if they had been its authors.

More frequent sources of error are subcontractors who are generally competent, but who lack specific insight into the needs of a local assessment. The remedy for this is close cooperation between assessment staff and contractor.

Effective decision-making procedures. Ineffective decision making could be an amorphous malaise that is observable only in late reports or contradictory public statements. However, if evaluators can be specific about decision-making difficulties, there may be clear implications for program improvement. Here are some things to look for: Are all decisions reserved for a boss who's out-of-town 50 percent of the time? Are decisions made by default? Are key decisions made by uninformed sources outside of the assessment staff? Are members of the staff satisfied with the way decisions are made?

Equal opportunity. Assessments may have special problems as equal-opportunity employers. Assessment is a "math-related" activity. Women have been taught to avoid math and may remove themselves from the competition for assessment jobs. If some minority group members had poor elementary and secondary training in mathematics, they may have eschewed math-related subjects in college. Special recruitment may be necessary to interrupt this trend.

Redress of grievances. Regular procedures should be established for hearing and redressing complaints against the assessment. If particular school or district results are in error, new reports should be printed.

Assessment staff are likely to receive a certain amount of hate-mail and irate phone calls. In the interest of fairness, responses to complaints should make clear exactly what will be done to correct the problem. In many instances, the appropriate answer is that no change will be made. Reasons for this staff stance should be clearly delineated. The assessment staff should only be faulted if they make false promises.

Timeliness. This criterion needs little elaboration. Good information is not helpful if it arrives after decisions have already been made.

When assessment results are collected for long-term census monitoring, delays of six months in reporting are not crucial. When assessment results are meant to be used for program evaluation or state-level decision-making, delays may render the reports useless. This item certainly overlaps with the effects considered in the next section.

Intended and Unintended Effects

People and Groups Who May be Affected

Earlier, I suggested some strategies for seeking the effects of assessment. Effects may be subtle and indirect. The evaluator must be a bit of a sleuth to detect them. For example, the true effect of the New York State Regents' Exam may not be to provide evidence for admission to college but to increase the amount of studying done by high school juniors and seniors.

Possible effects might be discovered by surveying diverse audiences. Then, purported effects must be tracked down and verified. The evaluator must be particularly cautious not to identify consumers who are only in the educational community. Educators are the more visible consumers and are perhaps better informed about assessment but important recipients exist in the larger community.

The Assessment Checklist identifies individuals and groups who may be affected by assessment:

Students
Teachers
Parents
Principals
District superintendents
District curriculum specialists
Regional personnel
State curriculum experts
State legislators
State governors
Taxpayers
Employers
Special interest groups
Researchers

The evaluator may add to the list. Each audience must then be studied in light of the possible effects.

This checkpoint incorporates process effects as well as outcomes. If children enjoy taking a test or if teachers feel threatened by the results of the test, these are effects of the assessment. The discussion of effects subsumes Stufflebeam's utility criteria of Relevance, Importance, Credibility, and Pervasiveness. The impact of the assessment is measured in terms of who uses the results and how widespread the effect is. Suppose one or two teachers testify that they have used assessment results to modify their classroom instruction. If they are isolated examples of this practice,

there is evidence of potential use but not of significant current use.

Scriven's discussion of Educational Significance suggests that at some point we go back to the original need. Is this program serving the purposes for which it was intended? Of course, we must also be on the lookout for unintended effects and cancelled purposes.

Kinds of Effects

The second set of entries in the Effects category contains some possible consequences of assessment to consider.

Outcome and process effects. In other evaluation contexts, process variables are usually given separate consideration. In this situation, however, processes are those described in the management portion of this paper. Things that happen to teachers and children in the process of test administration are actually side effects of the assessment.

Immediate and long-term effects. Immediate consequences of the assessment should be observable in program changes, public reactions, resource allocations, and development of new curricula. Long-term effects can be only estimated by the evaluator. Educational assessment is in its infancy. Some uses which may emerge in the future have not been identified. Reliance on the on-going nature of the information has not been established. The evaluator may have to be an advocate in trying to project the assessment's ultimate usefulness. What uses have emerged for data from analogous enterprises, the U. S. Census and the GNP, that were not thought of when the data collection began? Of course, the evaluator's advocacy should not subvert his responsibility to obtain confirming opinions of his projections or to report contrary views of the future as well.

Opinions. Attitudes toward testing and toward education may be influenced by an assessment program. For example, if teachers believe that an assessment program is unfair, they may loose confidence in all tests and convey this misgiving to their children. A positive outcome may occur when children have an opportunity to take "easy" tests. The California first-grade test is not meant to discriminate among children. It is used for baseline data to distinguish programs. Tests on which many children get a perfect score can still show differences between schools and programs. The California test consisted of questions that 80 or 90 percent of the children answered correctly. On the hardest question, half of the children got the right answer. Teachers reported that children thought the test was fun.

The affective component of accountability should also be considered under this heading. Public confidence in education may result simply because people have access to information.

Decisions, laws, resource allocation. Decisions are tangible, e.g., a new emphasis on math, government funding for consumer education, legislative requirements for minimum competency testing. The question is, which decisions are attributable to assessment results? One might begin by asking decision-makers, superintendents, and legislators. But many effects are more subtle. Scriven's modus operandi method cited previously may help us uncover more grass-roots influences. Are citizens voting against school bond issues because they won't pay higher taxes or have they lost confidence in the public schools? What evidence do citizens cite for or against public schools? How many editorials in selected newspapers refer to assessment results? How do interest groups use assessment to lobby for their purposes? The evaluator should review legislative debates on educational measures. Are

assessment results quoted? Decisions about resource allocation should be the easiest to document. What are the formulae for distributing federal and state funds? Within districts, how are special resources apportioned to schools?

There is probably no harm in assessment staffs collecting evidence that verifies the uses of assessment results as long as they do not throw out evidence of misuses. True, this may give a biased view since many isolated examples of use may not reflect pervasive use. However, it will be up to the evaluator to substantiate the extent of each type of use.

In searching for examples of information use, evaluators should be wary of the "common knowledge" fallacy. In the State Senate evaluation of the Minnesota Assessment (1975), there is this complaint:

> It is no secret that students from poor families do not perform as well as students from middle and upper income level families; just as it's no secret that boys read better than girls [this is backwards] and that students from homes exposed to reading materials outperform students from homes with no books around. (P.3)

Have such things always been known? What these authors claim is no secret may not have always been common knowledge. We ought to find out how much of this was publicly acknowledged before Coleman. What kinds of evidence were used to argue for ESEA Title I and Headstart? The Senate researchers may have a valid point if these are countless assessments all producing duplicate information; but that is a separate issue from whether the information is useful in itself.

Effects of assessment technology. Assessment efforts may produce expertise as well as data. Through numerous conferences with state and local educators NAEP has disseminated its methodological expertise and fostered sharing among other assessors. Shared expertise should improve the overall quality of assessment. However, evaluators should watch for instances where the imported model is inappropriate for local purposes. State depart-

ments may, in turn, give assistance to local districts who wish to conduct assessments for their own purposes. "Piggy-backing" is the term that has been adopted in the trade to signify local assessments that take advantage of the instrument development and sampling plan of the larger assessment organization. By oversampling in districts who ask for more information, state assessments can provide them with test data much more economically than if they had tried to conduct the assessment themselves.

Costs

The evaluator should collect cost data as carefully as he gathers effects data. Costs must include expenditures in dollars and time. It is a matter of style, however, whether one wishes to count emotional costs in this category or consider them as effects in checkpoint four.

Scriven (1974) made several requirements of cost data: "There should be some consideration of opportunity costs...What else could the district or state have done with the funds... Cost estimates and real costs...should be verified independently...[and], costs must, of course, be provided for the critical competitors". (p. 20-21)

In an assessment context, costs should include obvious things like the salaries of assessment staff and the cost of their offices. The most obvious cost is that of consultants or subcontractors since they present a bill. Costs which are often overlooked are the salaries of teachers and curriculum experts who are released by their districts to consult on objectives development or item writing. Also forgotten are the costs in teacher and student time while the assessment is underway. How many total hours are spent by district personnel and principals distributing and

collecting test materials. Be sure to count the time they spend on the telephone answering questions.

Synthesizing Checklist Findings

The checklist produces a collection of judgments--B+ on the sampling plan, C- on public knowledge about results. But, these are raw judgments. In the end, the evaluator must help combine the information so that the assessment can be judged. Unfortunately, a simple grade-point average will not do. Some of the checklist items are more important than others. Some must be combined or weighed one against the other.

The single most important checklist category is Effects. Technical aspects and management are important only because they serve the effects. Goals do not count for much if they are not accomplished.

Cost-benefit analysis will certainly be a part of the synthesis. This analytical procedure is borrowed from economics, but is not well refined since it is more difficult to assign dollar values to informed decisions or data preserved for future comparison than it is to compute the cost of capital equipment or increased tax revenues from 10,000 new jobs.
The most straightforward device is a tally sheet. What are the positive effects of the assessment? The evaluator should try to specify how big each effect is in terms of the number of school districts it reaches or the number of classroom teachers actually using the information. Opposite these entries are the costs and negative side effects. For example, the usefulness of the data to teachers will have to be weighed against a portion of the total assessment costs and against the number of teachers who were required to administer the tests but found no use for results.

Finally, the total assessment package must be judged in comparison to what Scriven calls Critical Competitors. How many of the assessment purposes would go unserved if the assessment program were discontinued? Might all or part of the assessment effects be accomplished by commercial publishers or by local districts. If a state assessment has also presumed to conduct classroom testing for its local districts, what are the gains in teacher appreciation of the data and uniformity. Are these gains worth the increased testing time and loss of local autonomy in selecting tests?

Although technical aspects of the assessment and management considerations may not be the focus of summative judgments about the assessment, these checkpoints are the ones more likely to provide diagnosis of needed improvements in the assessment. If a large-scale assessment is engaged in every-pupil testing but is providing redundant information to the classroom teacher, the evaluators should recommend sampling.

References

Airasian, P. W., & Madaus, G. F. A study of the sensitivity of school and program effectiveness measures (Report submitted to the Carnegie Corporation of New York). Unpublished manuscript, July, 1976.

Alkin, M. C. Evaluation theory development. Evaluation Comment, 1969, 2(1), 2-7. Reprinted in B. R. Worthen & J. R. Sanders (Eds.), Educational evaluation: Theory and practice. Worthington, Ohio: Charles A. Jones, 1973.

An evaluation of the National Assessment of Educational Progress, by the Site-Visit Team established by the National Center for Educational Statistics. Prepared under the auspices of the Evaluation Research Center, 1974.

An evaluation of the National Assessment of Educational Progress, by the Site-Visit Team established by the National Center for Educational Statistics, June 1975.

Coleman, J. S. Equality of educational opportunity. Washington, D.C.: U.S. Department of Health, Education, and Welfare, Office of Education, 1966.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Glass, G. V. Teacher effectiveness. In H. J. Walberg (Ed.), Evaluating educational performance. Berkeley: McCutchan, 1974.

Greenbaum, W. Measuring educational progress. New York: McGraw Hill, in press.

Hammond, R. L. Evaluation at the local level. Tucson, AZ: EPIC Evaluation Center (undated mimeo). Reprinted in B. R. Worthen & J. R. Sanders (Eds.), Educational evaluation: Theory and practice. Worthington, OH: Charles A. Jones, 1973.

House, E., Rivers, W., & Stufflebeam, D. An assessment of the Michigan accountability system (Under contract with the Michigan education Association and the National Education Association). March. 1974. Reprinted in the Evaluation Report Series, no. 2, Kalamazoo, MI: The Evaluation Center, Western Michigan University, 1976.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Minnesota educational assessment. St. Paul, MN: Office of Legislative Research, State of Minnesota, 1975.

Murphy, J. T., & Cohen, D. K. Accountability in education--the Michigan experience. The Public Interest, Summer, 1974, 36, 53-81.

Provus, M. M. Evaluation of ongoing programs in the public school system. In R. W. Tyler (Ed.), Educational evaluation: New roles, new means. The 68th Yearbook of the National Society for the Study of Education, Part II. Chicago: National Society for the Study of Education, 1969.

Rakow, E. A., Airasian, P. W., & Madaus, G. F. Assessing school and program effectiveness: Estimating hidden teacher effects. Paper read at annual meeting of the National Council on Measurement in Education, San Francisco, CA, April, 1976.

Response to the senate research report. St. Paul, MN: Office of Statewide Assessment, Minnesota Department of Education, 1975.

Sanders, J. R., & Nafziger, P. H. A basis for determining the adequacy of evaluation designs. Occasional Paper Series, no. 6, Kalamazoo, MI: The Evaluation Center, Western Michigan University, 1976.

Scriven, M. The methodology of evaluation. In R. E. Stake (Ed.), Curriculum evaluation. American Educational Research Association monograph series on evaluation, no. 1, Chicago: Rand McNally, 1967. Reprinted with revisions in B. R. Worthen & J. R. Sanders (Eds.), Educational evaluation: Theory and practice. Worthington, OH: Charles A. Jones, 1973.

Scriven, M. Evaluation perspectives and procedures. In W. J. Popham (Ed.), Evaluation in education. Berkeley: McCutchan, 1974.

Scriven, M. Evaluation bias and its control. Occasional Paper Series, no. 4, Kalamazoo, MI: The Evaluation Center, Western Michigan University, 1975.

Shepard, L. Reporting the results of statewide assessment. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1975.

Staff report to the Committee on Education, Florida House of Representatives, Richard S. Hodes, Chairman. Legislative concerns of Florida public school districts. February, 1976.

Staff response to the report: An assessment of the Michigan accountability system. Lansing, MI: Michigan Department of Education, May, 1974. Reprinted in the Evaluation Report Series, no. 2, Kalamazoo, MI: The Evaluation Center, Western Michigan University, 1976.

Stake, R. E. The countenance of educational evaluation. Teachers College Record, 1967, 68, 523-540. Reprinted in B. R. Worthen & J. R. Sanders (Eds.), Educational evaluation: Theory and practice. Worthington, OH: Charles A. Jones, 1973.

Stake, R. E. Evaluation design, instrumentation, data collection, and analysis of data. In J. L. David (Ed.), Educational evaluation. Columbus, OH: State Superintendent of Public Instruction, 1969. Reprinted with revisions in B. R. Worthen & J. R. Sanders (Eds.), Educational evaluation: Theory and practice. Worthington, OH: Charles A. Jones, 1973.

Stake, R. E. Program evaluation, particularly responsive evaluation. Paper presented at a conference on "New Trends in Evaluation," Goteborg, Sweden, October, 1973. Reprinted in Occasional Paper Series, no. 5, Kalamazoo, MI: The Evaluation Center, Western Michigan University, 1976.

Statewide pupil testing in New York. Office of Education Performance Review, in press.

Stufflebeam, D. L. Meta evaluation. Occasional Paper Series, no. 3, Kalamazoo, MI: The Evaluation Center, Western Michigan University, 1974. (a)

Stufflebeam, D. L. A response to the Michigan Education Department's defense of their accountability system. Occasional Paper Series, no. 1, Kalamazoo, MI: The Evaluation Center, Western Michigan University, 1974. (b)

Stufflebeam, D. L., Foley, J. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Merriman, H. O., & Provus, M. M. Educational evaluation and decision making in education. Itasca, IL: Peacock, 1971.

Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-276.

Teacher's manual, reading test (second and third grades). Sacramento, CA: California Assessment Program, California Department of Education, 1974.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. Unobtrusive measures. Chicago: Rand McNally, 1966.

Webster, J., Millman, J., & Gordon, E. Statewide pupil testing in New York state (a report by the Blue Ribbon Panel). Albany, NY: Office of Education Performance Review, December, 1974.

Woodward, R., & Bernstein, C. The final days. New York: Simon & Schuster, 1976.

Womer, F., & Lehmann, I. Reactions to the Oregon statewide assessment program (three reports). 1974.

Additional copies of papers in this series are available upon request for a fee of $2.00. Previous papers in this series are listed below. Please direct your inquiries to

Mary Anne Bunda, Series Editor
The Evaluation Center
College of Education
Western Michigan University
Kalamazoo, Michigan 49008

1. A Response to the Michigan Education Department's Defense of Their Accountability System by Daniel L. Stufflebeam.
2. An Administrator's Perspective of Evaluation by Jack P. Taylor.
3. Meta-Evaluation by Daniel L. Stufflebeam.
4. Evaluation Bias and Its Control by Michael Scriven.
5. Program Evaluation, Particularly Responsive Evaluation by Robert E. Stake.
6. A Basis for Determining the Adequacy of Evaluation Designs by James R. Sanders and Dean H. Nafziger.
7. Needed: Instruments as Good as Our Eyes by Henry M. Brickell.
8. Assessing the Impact of Planned Social Change by Donald T. Campbell.
9. A Checklist for Evaluating Large-Scale Assessment Programs by Lorrie A. Shepard.
10. Standards and Criteria by Gene V Glass.
11. Performance Standards by Nancy W. Burton.